

# Blind Compressed Sensing Over a Structured Union of Subspaces

Jorge Silva, *Member, IEEE*, Minhua Chen, Yonina C. Eldar, *Senior Member, IEEE*, Guillermo Sapiro, *Senior Member, IEEE*, and Lawrence Carin, *Fellow, IEEE*

**Abstract**—This paper addresses the problem of simultaneous signal recovery and dictionary learning based on compressive measurements. Multiple signals are analyzed jointly, with multiple sensing matrices, under the assumption that the unknown signals come from a union of a small number of disjoint subspaces. This problem is important, for instance, in image inpainting applications, in which the multiple signals are constituted by (incomplete) image patches taken from the overall image. This work extends standard dictionary learning and block-sparse dictionary optimization, by considering compressive measurements (e.g., incomplete data). Previous work on blind compressed sensing is also generalized by using multiple sensing matrices and relaxing some of the restrictions on the learned dictionary. Drawing on results developed in the context of matrix completion, it is proven that both the dictionary and signals can be recovered with high probability from compressed measurements. The solution is unique up to block permutations and invertible linear transformations of the dictionary atoms. The recovery is contingent on the number of measurements per signal and the number of signals being sufficiently large; bounds are derived for these quantities. In addition, this paper presents a computationally practical algorithm that performs dictionary learning and signal recovery, and establishes conditions for its convergence to a local optimum. Experimental results for image inpainting demonstrate the capabilities of the method.

## I. INTRODUCTION

The problem of learning a dictionary for a set of signals has received considerable attention in recent years. This problem is known to be ill-posed in general, unless constraints are imposed on the dictionary and signals. One such constraint is the assumption that the signals  $\mathbf{x}_i \in \mathbb{R}^n, i = 1, \dots, N$ , can be sparsely represented under an unknown dictionary, *i.e.*, each vector  $\mathbf{x}_i$  can be written as

$$\mathbf{x}_i = \mathbf{D}\mathbf{s}_i + \boldsymbol{\epsilon}_i, \quad (1)$$

where  $\mathbf{D} \in \mathbb{R}^{n \times r}$  is the dictionary, and  $\mathbf{s}_i \in \mathbb{R}^r$  are sparse coefficient vectors satisfying  $\|\mathbf{s}_i\|_0 \ll r$ . The residual energy

$\|\boldsymbol{\epsilon}_i\|_2^2$  is assumed to be small. For example, [1] derives conditions to ensure the uniqueness of  $\mathbf{D}$  and the representations  $\mathbf{s}_i$ , for the case in which all the coordinates of the signals  $\mathbf{x}_i$  are observed. In this setting, the problem is known as dictionary learning (DL) or, sometimes, *collaborative* DL to emphasize the fact that multiple signals are considered. Note that, in DL and in most related literature, “uniqueness” is defined up to an equivalence class involving permutations and rotations of the dictionary atoms; we follow the same convention throughout this paper. In the DL setting, several algorithms have been proposed for estimating  $\mathbf{D}$  and  $\mathbf{s}_i$ . Examples include K-SVD [1] and the method of optimal directions (MOD) [18], both of which enjoy local convergence properties. More recently, [16] has derived local convergence guarantees for  $\ell_1$  minimization applied to DL when the signals are sparse.

A more structured type of sparsity is considered in block-sparse dictionary optimization (BSDO) [26], in which it is assumed that the nonzero coordinates of  $\mathbf{s}_i$  (active atoms) occur in blocks rather than in arbitrary locations. This property is called block sparsity [12], and is important for the analysis of signals that belong to a union of a small number of subspaces, as described in [13]. The standard BSDO framework, like DL, assumes that all coordinates of the signals  $\mathbf{x}_i$  are observed. The block structure, *i.e.*, the number of atoms and composition of each block of the dictionary, is in general not known *a priori* and should be estimated as part of the DL process. The BSDO algorithm reduces to K-SVD when the block size is one (standard sparsity).

While, technically, the set of all  $k$ -sparse signals in  $\mathbb{R}^r$  is itself a union of  $\binom{r}{k}$  subspaces, it greatly simplifies the problem if one considers the more structured setting of block sparsity. This reduces the number of subspaces to  $\binom{L}{K}$ , where  $L$  is the total number of blocks in the dictionary and  $K$  is the number of blocks that are active (block sparsity reduces to standard sparsity when all blocks are singletons). Block sparsity is closely related to the group LASSO in statistics literature [20], and also to the mixture of factor analyzers (MFA) statistical model, as noted in [8]. The particular case for which only one block is active is called one-block sparsity and corresponds to a union of  $\binom{L}{1} = L$  possible subspaces, which is a dramatic simplification compared to the unstructured sparsity case. While this might at first sound limiting, the authors in [30] have obtained state-of-the-art image restoration results with a one-block sparsity model, and a variety of alternative methods have been proposed for this special case [14], [29], [30].

In applications we often do not have access to data  $\mathbf{x}_i$ , but

J. Silva is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: jg.silva@duke.edu).

M. Chen is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: minhua.chen@ee.duke.edu).

Y. C. Eldar is with the Department Electrical Engineering, The Technion-Israel Institute of Technology, Haifa 32000, Israel (e-mail: yonina@ee.technion.ac.il).

G. Sapiro is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: guille@umn.edu).

L. Carin is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: lcarin@ee.duke.edu).

only to compressive measurements

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i = \mathbf{A}\mathbf{D}\mathbf{s}_i, \quad (2)$$

where  $\mathbf{A} \in \mathbb{R}^{m \times n}$  has  $m < n$  rows, so that  $\mathbf{y}_i \in \mathbb{R}^m$ . In order to enable recovery of  $\mathbf{s}_i$  from  $\mathbf{y}_i$  even when  $\mathbf{D}$  is known, the sensing matrix  $\mathbf{A}$  must satisfy incoherence properties with respect to  $\mathbf{D}$ , as prescribed by compressed sensing (CS) theory [6], [9]. Learning  $\mathbf{D}$  from compressive measurements is called blind compressed sensing (blind CS), and does not admit a unique solution unless additional structure is imposed on  $\mathbf{D}$  [17]. See also the approach in [10] for simultaneous dictionary learning and sensing matrix design in the CS scenario.

In this paper, we address simultaneous estimation of one-block-sparse signals and the corresponding dictionary, given only compressive measurements. This unifies blind CS and BSDO (for the one-block-sparse case in particular). It is known that the standard blind CS problem, where a fixed sensing matrix  $\mathbf{A}$  is used, does not admit a unique solution in general, although a number of special cases of dictionaries for which such a solution exists have been identified in [17]. These special cases are: (i) finite sets of bases (*i.e.*, it is known that the dictionary is one member of a finite set of known bases for  $\mathbb{R}^n$ ); (ii) sparse dictionaries (*i.e.*, the atoms of the dictionary themselves admit sparse representations) and (iii) block-diagonal dictionaries, with each block composed of orthogonal columns.

We are motivated in part by the problem of inpainting and interpolating an image [2], [31], where one observes an incomplete image, *i.e.*, we only know the intensity values at a subset of pixel locations (or a subset of their linear combinations). Additionally, the image is processed in (often overlapping) patches, which we convert to  $n$ -dimensional vectors (our signals of interest). We observe  $m_i < n$  pixels in each patch, indexed by  $i$ , with these  $m_i$  pixels selected at random. Therefore, the locations of the missing pixels are in general (at least partially) different for each patch. This means that, unlike the classical blind CS setting, the sensing matrix is not the same for all signals; we denote the sensing matrix for patch  $i$  as  $\mathbf{A}_i \in \mathbb{R}^{m_i \times n}$ . The assumption of multiple and distinct  $\mathbf{A}_i$  is crucial for solving the problem of interest, as it enables the use of existing results from matrix completion [24]. Moreover, it has been demonstrated in [29] that using multiple  $\mathbf{A}_i$  improves inpainting and interpolation performance.

In conventional image inpainting, each  $\mathbf{A}_i$  consists of a randomly chosen subset of  $m_i$  rows of the  $n \times n$  identity matrix (thereby selecting  $m_i$  pixels). Successful estimation of the missing pixel intensity values is contingent on each patch being representable in terms of a small subset of the columns in the dictionary  $\mathbf{D}$  [31]. In other words, there needs to exist some  $\mathbf{D}$  such that the  $\mathbf{s}_i$  are (at least approximately) sparse. However, our analysis is not restricted to  $\mathbf{A}_i$  being defined as in conventional inpainting. Other constructions typically used in CS, such as matrices with i.i.d. random-subgaussian [27] entries, or with rows drawn uniformly at random from an orthobasis [6], can be employed. Moreover, in many settings it is appropriate to assume that the patches belong to a union of disjoint subspaces [23], [30], with the number of subspaces being small. This motivates the one-block-sparsity assumption

underlying our theoretical and computational results. This assumption, coupled with the use of multiple  $\mathbf{A}_i$ , allows us to ensure recovery of  $\mathbf{D}$  and the  $\mathbf{s}_i$  under milder conditions on  $\mathbf{D}$  and on the number of vectors  $\mathbf{y}_i$ , as compared to standard blind CS.

We show that unique recovery of the dictionary and the one-block-sparse signals is guaranteed with high probability, albeit with high computational effort, if the number of measurements per signal and the number of signals are sufficiently large. We derive algorithm-independent bounds for these quantities, thereby extending DL by considering compressive measurements and establishing a connection with matrix-completion theory. Our results reduce to those known for DL when the signals are fully observed.

Additionally, we present a computationally feasible algorithm that performs DL and recovery of block-sparse signals based on compressive measurements with multiple sensing matrices. We automatically learn the block structure, in the same way as the BSDO algorithm [26]; the size (number of atoms) and composition of each block is not known *a priori* and we only need to specify a maximum block size. Our algorithm is closely related to BSDO, the main differences being that one-block sparsity and compressive measurements are considered. The estimates of  $\mathbf{D}$ ,  $\mathbf{s}_i$  and the block structure are found by alternating least-squares minimization.

It is shown that the algorithm converges to a local optimum under mild conditions, which we derive. This convergence analysis is not available for most other one-block-sparsity promoting methods, such as [8], or for most methods that rely on standard sparsity, such as [31]. An exception is the analysis in [16] pertaining to  $\ell_1$  minimization, where local convergence is proven when the number of measurements is at least  $O(n^3k)$  for signals with sparsity level  $k$ . However, compressive measurements are not considered. Besides our global uniqueness result, our method provably attains local convergence for, at most,  $O(nk \log n)$  measurements, although our one-block sparsity assumption is stronger than in [16]. The approach proposed in [30] also converges, but involves an even stronger assumption akin to intra-block sparsity, and does not include an algorithm-independent uniqueness analysis. Compelling experimental results are presented below, demonstrating the ability of our algorithm to perform inpainting of real images with a significant fraction of missing pixels.

The remainder of the paper is organized as follows. Section II presents preliminary definitions and a formulation of the optimization problem. Our uniqueness result is presented in Section III. Sections IV and V respectively describe the proposed algorithm and the corresponding proof of convergence to a local optimum. Experimental results are described in Section VI and concluding remarks are given in Section VII.

## II. PROBLEM FORMULATION

### A. Preliminaries

Assume vectors  $\mathbf{y}_i \in \mathbb{R}^{m_i}$ ,  $i = 1, \dots, N$  are observed, such that

$$\mathbf{y}_i = \mathbf{A}_i \mathbf{x}_i = \mathbf{A}_i (\mathbf{D} \mathbf{s}_i + \epsilon_i) \quad (3)$$

where  $\mathbf{x}_i \in \mathbb{R}^n$  is an unknown signal,  $\mathbf{A}_i \in \mathbb{R}^{m_i \times n}$  is a known sensing matrix,  $\mathbf{D} \in \mathbb{R}^{n \times r}$  is an unknown dictionary,  $\mathbf{s}_i \in \mathbb{R}^r$  is an unknown sparse vector of coefficients such that  $\mathbf{x}_i = \mathbf{D}\mathbf{s}_i + \epsilon_i$ , with small residual  $\|\epsilon_i\|_2$ . We focus on the noiseless case, although our analysis can be extended straightforwardly to include observation noise, following [4]. Given  $\mathbf{y}_i$  we would like to estimate  $\mathbf{x}_i$  by finding  $\mathbf{D}$  and  $\mathbf{s}_i$ . Thus, we are interested in solving

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{s}_1, \dots, \mathbf{s}_N} \sum_{i=1}^N \|\mathbf{y}_i - \mathbf{A}_i \mathbf{D} \mathbf{s}_i\|_2^2, \\ \text{s.t. } \mathbf{s}_i \text{ one-block sparse} \end{aligned} \quad (4)$$

with our estimates denoted  $\hat{\mathbf{D}}, \hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_N$ .

It is assumed that each  $\mathbf{x}_i$  lives in a subspace specified by a subset of the columns of  $\mathbf{D}$ , so that there exists a permutation of the columns such that  $\mathbf{x}_i = \mathbf{D}\mathbf{s}_i$  and the coefficient vector  $\mathbf{s}_i$  is one-block sparse, as defined below. It is also assumed that  $\mathbf{D}$  is composed of blocks (subsets of columns) corresponding to the blocks of  $\mathbf{s}_i$ , with the cardinality of the blocks summing to  $r$ . The cardinality and composition of each block are unknown, although we fix a maximum cardinality as in BSDO. The atoms in each block are assumed orthonormal, in order to avoid scaling indeterminacies and also to obviate concerns with sub-coherence (see [12] for a discussion). The expression in (4), particularly the  $\ell_2$  norm, is equivalent to a maximum-likelihood solution for the unknown model parameters, assuming that the components of  $\epsilon_i$  are i.i.d. Gaussian and the model deviation is negligible.

**Definition 1 (Block sparsity and one-block sparsity).** Let the dictionary  $\mathbf{D} \in \mathbb{R}^{n \times r}$  have a block structure such that  $\mathbf{D} = [\mathbf{D}[1] \cdots \mathbf{D}[L]]$ , where each  $\mathbf{D}[\ell], \ell \in \{1, \dots, L\}$  is a unique subset of the columns in  $\mathbf{D}$ , and the columns of  $\mathbf{D}[\ell]$  are assumed orthonormal for all  $\ell$ . Following [12], a signal  $\mathbf{s}_i$  is  $K$ -block-sparse under dictionary  $\mathbf{D}$  if it admits a corresponding block structure  $\mathbf{s}_i = [\mathbf{s}_i[1]^T \cdots \mathbf{s}_i[L]^T]^T$  and if  $\mathbf{s}_i$  has zero entries everywhere except at a subset  $\mathbf{s}_i[\ell_1], \dots, \mathbf{s}_i[\ell_K]$  of size  $K \ll L$ , corresponding to blocks  $\mathbf{D}[\ell_1], \dots, \mathbf{D}[\ell_K]$  of  $\mathbf{D}$ . Each dictionary block  $\mathbf{D}[\ell] \in \mathbb{R}^{n \times k_\ell}$  and each coefficient block  $\mathbf{s}_i[\ell] \in \mathbb{R}^{k_\ell}$ . In general, the block size  $k_\ell$  is not known *a priori*, although it is common to define a maximum size  $k_{\max}$ .

We are specifically interested in one-block sparsity, where  $K = 1$ . An interpretation of one-block sparsity is that it corresponds to signals that live in a union of  $L$  linear subspaces, with each subspace spanned by the columns of one block  $\mathbf{D}[\ell]$ . If block  $\mathbf{D}[\ell]$  has  $k_\ell$  columns, then its subspace is of dimension  $k_\ell$ . Figure 1 illustrates the concept of one-block sparsity. The index set of all signals which use block  $\ell$  is  $\omega_\ell = \{i : \mathbf{s}_i[\ell] \neq \mathbf{0}\}$ . Note that for one-block-sparse signals there can be no overlap between  $\omega_{\ell_1}$  and  $\omega_{\ell_2}$  for  $\ell_1 \neq \ell_2$ .

The optimization problem (4) can be decomposed as

$$\begin{aligned} \min_{\omega_\ell, \mathbf{D}[\ell], \mathbf{s}_1[\ell], \dots, \mathbf{s}_N[\ell]} \sum_{i \in \omega_\ell} \|\mathbf{y}_i - \mathbf{A}_i \mathbf{D}[\ell] \mathbf{s}_i[\ell]\|_2^2, \\ \text{s.t. } \mathbf{D}[\ell]^T \mathbf{D}[\ell] = \mathbf{I}, \end{aligned} \quad (5)$$

for all  $\ell$ , where  $\mathbf{s}_i[\ell] \in \mathbb{R}^{k_\ell}$  is the  $\ell$ -th block of  $\mathbf{s}_i$ . The solution to problem (4) is only unique up to the following equivalence

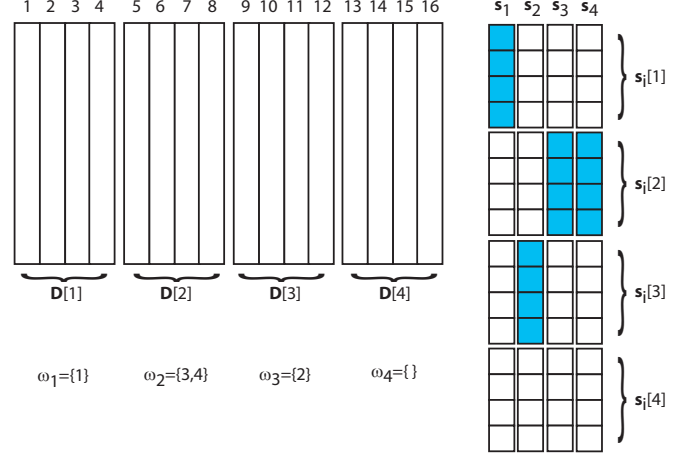


Fig. 1. Illustration of the concept of one-block sparsity. The figure represents a dictionary with  $r = 16$  columns, arranged consecutively in blocks  $\mathbf{D}[1], \dots, \mathbf{D}[4]$ , all of size four. Also represented is a coefficient matrix with columns  $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$  and  $\mathbf{s}_4$ . Non-zero coefficient values are colored, while zero-valued coefficient values are blank. All signals are one-block sparse, as they only use one block each. Note that there exists no overlap between index sets  $\omega_1, \omega_2, \omega_3$  and  $\omega_4$ .

class.

**Definition 2 (Equivalence class for  $\hat{\mathbf{D}}, \hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_N$ ).** Given a solution  $(\hat{\mathbf{D}}, \hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_N)$  to (4), with  $\hat{\mathbf{D}} = [\hat{\mathbf{D}}[1] \cdots \hat{\mathbf{D}}[L]]$  and  $\hat{\mathbf{s}}_i = [\hat{\mathbf{s}}_i[1]^T \cdots \hat{\mathbf{s}}_i[L]^T]^T, i = 1, \dots, N$ , any permutation  $\ell'_1, \dots, \ell'_L$  of the indices  $1, \dots, L$  corresponds to an equivalent solution  $(\hat{\mathbf{D}}', \hat{\mathbf{s}}'_1, \dots, \hat{\mathbf{s}}'_N)$  of the form  $\hat{\mathbf{D}}' = [\hat{\mathbf{D}}[\ell'_1] \cdots \hat{\mathbf{D}}[\ell'_L]]$  and  $\hat{\mathbf{s}}'_i = [\hat{\mathbf{s}}_i[\ell'_1]^T \cdots \hat{\mathbf{s}}_i[\ell'_L]^T]^T, i = 1, \dots, N$ . Additionally, any sequence of invertible matrices  $\mathbf{T}_\ell \in \mathbb{R}^{k_\ell \times k_\ell}, \ell = 1, \dots, L$ , where  $k_\ell$  is the size of the  $\ell$ -th block, also corresponds to an equivalent solution of the form  $\hat{\mathbf{D}}' = [\hat{\mathbf{D}}'[1] \cdots \hat{\mathbf{D}}'[L]]$  and  $\hat{\mathbf{s}}'_i = [(\hat{\mathbf{s}}_i[1])^T \cdots (\hat{\mathbf{s}}_i[L])^T]^T, i = 1, \dots, N$ , with  $\mathbf{D}[\ell]' = \mathbf{D}[\ell] \mathbf{T}_\ell$  and  $\mathbf{s}_i[\ell]' = \mathbf{T}_\ell^{-1} \mathbf{s}_i[\ell]$ .

Our first goal is to develop conditions under which (4) admits a unique solution. To this end we note that the index sets  $\omega_\ell$  are unknown, and therefore estimates  $\hat{\omega}_\ell$  must be found. This is due to the fact that, *a priori*, we do not know the correct assignments of columns to blocks. In our uniqueness analysis, we follow the same strategy as [1] and assume that all possible index sets can be tested. This is admittedly impractical in general, and it is done only for the purpose of constructing the uniqueness proof. The algorithm proposed in Section IV provides estimates of  $\omega_\ell$  without exhaustive search.

For a given  $\omega_\ell$ , (5) is an instance of the weighted orthonormal Procrustes problem (WOPP), which in general requires iterative optimization without guarantee of convergence to a global optimum [21]. In the following, we establish conditions under which uniqueness can be guaranteed. To this end we show that (5) can be written as a matrix-completion problem, and then rely on known results developed in that context.

## B. Formulation as a matrix completion problem

To reformulate (5), define  $\mathbf{X}_{\omega_\ell} \in \mathbb{R}^{n \times |\omega_\ell|}$  as the matrix whose columns are the signals  $\mathbf{x}_i$  that use block  $\ell$ , i.e., with  $i \in \omega_\ell$ , and similarly define  $\mathbf{S}_{\omega_\ell}[\ell] \in \mathbb{R}^{k_\ell \times |\omega_\ell|}$  as the matrix whose columns are the coefficient blocks  $\mathbf{s}_i[\ell]$  for  $i \in \omega_\ell$ .

Since  $\mathbf{D}[\ell]$  has only  $k_\ell$  columns and  $\mathbf{S}_{\omega_\ell}[\ell]$  has  $k_\ell$  rows, it is clear that  $\mathbf{X}_{\omega_\ell} = \mathbf{D}[\ell] \mathbf{S}_{\omega_\ell}[\ell]$  has rank at most  $k_\ell$ . In fact, unless the null space of  $\mathbf{D}[\ell]$  intersects the range of  $\mathbf{S}_{\omega_\ell}$ , the rank is exactly  $k_\ell$ . This should not happen in general, barring degeneracies that are precluded by the conditions assumed in our results. This suggests the strategy of treating each subproblem (5), assuming  $\omega_\ell$  is known and hence the subspace selection has been determined (in our method we address estimation of  $\omega_\ell$ ), as a low-rank matrix completion problem, as we illustrate next.

Define  $\tilde{\mathbf{A}} \in \mathbb{R}^{M \times n}$  with  $M \geq n$ , as the matrix constructed by taking the union of the unique rows of all sensing matrices  $\mathbf{A}_i$ , for  $i \in \omega_\ell$  (in the interest of notation brevity, we omit the dependence of  $\tilde{\mathbf{A}}$  on  $\ell$ ). For example, in the aforementioned inpainting problem, for which each  $\mathbf{A}_i$  is defined by selecting rows at random from the  $n \times n$  identity matrix, denoted  $\mathbf{I}_{n \times n}$ , we have  $\tilde{\mathbf{A}}$  equal to a subset of  $\mathbf{I}_{n \times n}$  and in the limit, given enough  $\mathbf{A}_i$ ,  $\tilde{\mathbf{A}} = \mathbf{I}_{n \times n}$  (up to row permutation). However, the discussion below considers more-general  $\mathbf{A}_i$ , for example composed in terms of draws from a subgaussian distribution. Let  $\mathbf{Y}_{\omega_\ell} \in \mathbb{R}^{M \times |\omega_\ell|}$  be defined as  $\mathbf{Y}_{\omega_\ell} = \tilde{\mathbf{A}} \mathbf{X}_{\omega_\ell}$ . When performing measurements, we do not observe all elements of  $\mathbf{Y}_{\omega_\ell}$ ; for each column  $i \in \omega_\ell$ , we only have access to the entries selected by  $\mathbf{A}_i$ , these corresponding to the matching observed vector  $\mathbf{y}_i$ . This is illustrated in Figure 2, which also shows a pictorial comparison with the cases of fully measured signals (DL) and a single measurement matrix  $\mathbf{A}$  (standard blind CS).

Denoting the locations of the observed entries of  $\mathbf{Y}_{\omega_\ell}$  as

$$\Omega = \{(u, v) : Y_{\omega_\ell}(u, v) \text{ is observed}\}, \quad (6)$$

the observation model can be written as

$$P_\Omega(\mathbf{Y}_{\omega_\ell}) = P_\Omega(\tilde{\mathbf{A}} \mathbf{X}_{\omega_\ell}) = P_\Omega(\tilde{\mathbf{A}} \mathbf{D}[\ell] \mathbf{S}_{\omega_\ell}[\ell]), \quad (7)$$

where  $P_\Omega(\cdot)$  is the operator that extracts the values of its argument at locations indexed by  $\Omega$ . Each subproblem (5) can then be reformulated as

$$\begin{aligned} \min_{\omega_\ell, \mathbf{D}[\ell], \mathbf{S}_{\omega_\ell}[\ell]} & \left\| P_\Omega(\mathbf{Y}_{\omega_\ell}) - P_\Omega(\tilde{\mathbf{A}} \mathbf{D}[\ell] \mathbf{S}_{\omega_\ell}[\ell]) \right\|_F \\ \text{s.t.} & \quad \mathbf{D}[\ell]^T \mathbf{D}[\ell] = \mathbf{I}, \end{aligned} \quad (8)$$

where  $\|\cdot\|_F$  is the Frobenius norm. If  $\mathbf{X}_{\omega_\ell}$  has rank  $k_\ell$  (the block size), assuming  $\tilde{\mathbf{A}}$  has rank greater or equal to  $k_\ell$  and does not reduce the rank of  $\mathbf{X}_{\omega_\ell}$  (since  $k_\ell$  is generally small and  $\tilde{\mathbf{A}}$  is assumed to have rank  $n$ , these conditions for  $\tilde{\mathbf{A}}$  are anticipated to be typical), we have that  $\mathbf{Y}_{\omega_\ell}$  is also of rank  $k_\ell$ . This establishes the fact that, when  $\mathbf{X}_{\omega_\ell}$  has low rank, so does  $\mathbf{Y}_{\omega_\ell}$  (matrix completion results are most useful when the rank is low, although they hold for any value of the rank). Each subproblem (8) can then be solved by first completing the matrix  $\mathbf{Y}_{\omega_\ell}$  and then obtaining  $\mathbf{X}_{\omega_\ell}$  and, subsequently,  $\mathbf{D}[\ell]$  and  $\mathbf{S}_{\omega_\ell}[\ell]$ . Note that, although this is a two-step process, if  $\tilde{\mathbf{A}}$  has rank  $n$  then it can be inverted to find a unique mapping from any estimated complete matrix  $\mathbf{Y}_{\omega_\ell}$  to a corresponding estimated  $\mathbf{X}_{\omega_\ell}$ . As long as  $\mathbf{X}_{\omega_\ell}$  can be correctly estimated, the solution to (8) can then be found by a singular value decomposition (SVD) of  $\mathbf{X}_{\omega_\ell}$ . This is due to the fact that, for  $\text{rank}(\mathbf{D}[\ell] \mathbf{S}_{\omega_\ell}[\ell]) = k_\ell$ , SVD minimizes the Frobenius norm

of the residual  $\mathbf{X}_{\omega_\ell} - \mathbf{D}[\ell] \mathbf{S}_{\omega_\ell}[\ell]$  w.r.t.  $\mathbf{D}[\ell]$  and  $\mathbf{S}_{\omega_\ell}[\ell]$ , by the Eckart-Young theorem [11].

Following matrix completion theory [5], [7], [24], [25], if  $\mathbf{Y}_{\omega_\ell}$  is truly of low rank (and additional technical conditions are met), it can be correctly completed with high probability by solving the convex program

$$\begin{aligned} \text{minimize} & \quad \|\hat{\mathbf{Y}}_{\omega_\ell}\|_* \\ \text{s.t.} & \quad P_\Omega(\hat{\mathbf{Y}}_{\omega_\ell}) = P_\Omega(\mathbf{Y}_{\omega_\ell}), \end{aligned} \quad (9)$$

where  $\|\cdot\|_*$  is the *nuclear norm*, which is defined for a generic matrix  $\mathbf{Z}$  of rank  $k$  as the sum of its singular values, *i.e.*,

$$\|\mathbf{Z}\|_* = \sum_{i=1}^k \gamma_i(\mathbf{Z}), \quad (10)$$

with  $\gamma_i(\mathbf{Z})$  indicating the  $i$ -th singular value of  $\mathbf{Z}$ . Importantly, problem (9) is *not* equivalent to (8), since a solution of (8) may not be a solution of (9). This is a reflection of the fact that there may exist a high-rank  $\hat{\mathbf{Y}}_{\omega_\ell}$  which solves (8) but has high nuclear norm. However, the opposite holds: a solution of (9) is always a solution of (8), and therefore of (5). Moreover, the solution of (9) will have low rank and therefore produce  $\mathbf{D}[\ell]$  with the smallest possible block size  $k_\ell$ , which is clearly desirable. Thus, the nuclear norm formulation yields those solutions of the original problem (8) that are *useful*.

In the following section, we state conditions on the number of observed entries and on  $\tilde{\mathbf{A}}$  for successful recovery of  $\mathbf{Y}_{\omega_\ell}$ ,  $\mathbf{X}_{\omega_\ell}$  and, subsequently, of  $\mathbf{D}[\ell]$  and  $\mathbf{s}_1[\ell], \dots, \mathbf{s}_N[\ell]$ , assuming exhaustive search over  $\omega_\ell$ , by exploiting the connection between (5), (8) and (9). In Section IV we propose an algorithm that does not require exhaustive search and is guaranteed to converge to a local minimum under mild conditions here established. The notation used in this paper is summarized in Table 1 below, for ease of reference.

TABLE I  
QUICK NOTATION GUIDE

$\mathbf{y}_i \in \mathbb{R}^{m_i}$	Observed vector $i$
$m_i$	Number of observed coordinates of vector $\mathbf{y}_i$
$\mathbf{x}_i \in \mathbb{R}^n$	Unknown signal $i$
$\mathbf{s}_i \in \mathbb{R}^r$	Sparse coefficient vector $i$
$\mathbf{D} \in \mathbb{R}^{n \times r}$	Dictionary
$\mathbf{D}[\ell] \in \mathbb{R}^{n \times k_\ell}$	Dictionary block $\ell$ , <i>i.e.</i> the $\ell$ -th subset of the columns of $\mathbf{D}$
$k_\ell$	Number of atoms in block $\ell$
$\omega_\ell = \{i : \mathbf{s}_i[\ell] \neq \mathbf{0}\}$	Set of indexes of the signals that use block $\ell$ of the dictionary $\mathbf{D}$ with coefficients $\mathbf{s}_i$
$ \omega_\ell $	Number of signals that use block $\ell$
$\mathbf{X}_{\omega_\ell} \in \mathbb{R}^{n \times  \omega_\ell }$	Subset of the signals associated with block $\ell$
$\mathbf{S}_{\omega_\ell} \in \mathbb{R}^{r \times  \omega_\ell }$	Subset of the coefficient vectors associated with block $\ell$
$\mathbf{S}_{\omega_\ell}[\ell] \in \mathbb{R}^{k_\ell \times  \omega_\ell }$	Block $\ell$ of $\mathbf{S}_{\omega_\ell}$ , <i>i.e.</i> the $\ell$ -th subset of the rows, corresponding to $\mathbf{D}[\ell]$
$\mathbf{s}_i[\ell]$	Block $\ell$ of sparse coefficient vector $\mathbf{s}_i$ ( $\ell$ -th subset of the rows)
$\mathbf{A}_i \in \mathbb{R}^{m_i \times n}$	Sensing matrix for vector $i$
$\tilde{\mathbf{A}} \in \mathbb{R}^{M \times n}$	Union of the rows of multiple sensing matrices
$\mathbf{Y}_{\omega_\ell} \in \mathbb{R}^{M \times  \omega_\ell }$	Incompletely observed data matrix associated with block $\ell$
$\Omega$	Observed locations of $\mathbf{Y}_{\omega_\ell}$
$\otimes$	Kronecker product

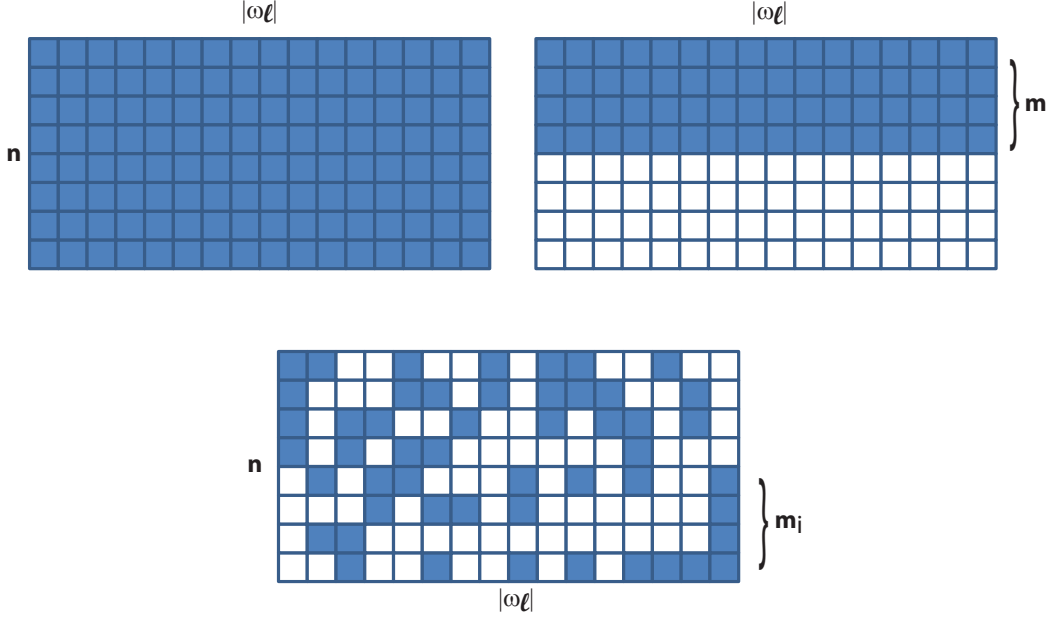


Fig. 2. Illustration of the observed elements of  $\mathbf{Y}_{\omega_\ell}$  with multiple measurement matrices (bottom), compared with the DL (top left) and standard blind CS (top right) cases. The colored squares represent observed elements, while blank squares represent unobserved elements of  $\mathbf{Y}_{\omega_\ell}$ . Here we assume that  $\mathbf{Y}_{\omega_\ell}$  has  $M = n$  rows.

### III. UNIQUENESS RESULT

Our results rely on the concepts of *spark* and *coherence*, which we now define. The spark  $\sigma(\cdot)$  of a matrix is the smallest number of columns that are linearly dependent; the number of linearly independent columns is the rank. Additionally, let  $\mathcal{U}$  be a subspace of  $\mathbb{R}^n$  with dimension  $k$  and spanned by vectors  $\{\mathbf{z}_u\}_{u=1,\dots,k}$ . The coherence of  $\mathcal{U}$  vis-à-vis the standard basis  $\{\mathbf{e}_v\}_{v=1,\dots,n}$  is defined as [24]

$$\mu(\mathcal{U}) = \frac{n}{k} \max_{u,v} \|\mathbf{z}_u^T \mathbf{e}_v\|^2. \quad (11)$$

This quantity is very similar to the coherence of a matrix, as defined in [28]. Our results also build on the following DL uniqueness conditions.

**Definition 3 (Uniqueness conditions for DL [1]).**

- **Support:**  $\|\mathbf{s}_i\|_0 = k < \frac{\sigma(\mathbf{D})}{2}, \forall i$
- **Richness:** There exist at least  $k+1$  signals for every possible combination of  $k$  atoms from  $\mathbf{D}$ . With regular sparsity level  $k$ , this amounts to at least  $(k+1)\binom{r}{k}$  signals. For one-block sparsity with  $L$  blocks of size  $k_\ell, \ell = 1, \dots, L$ , however, we need  $\sum_{\ell=1}^L (k_\ell+1)$  signals, which is typically a far smaller number.
- **Non-degeneracy:** Given  $k+1$  signals from the same combination of  $k$  atoms, their rank is exactly  $k$  (general position within the subspace). Similarly, any  $k+1$  signals from different combinations must have rank  $k+1$ .

These conditions apply to the case of fully measured signals and constitute a limiting case, since our problem reduces to DL when all  $\mathbf{A}_i$  are equal to the identity matrix. Even in this limiting case, however, the one-block sparsity condition has the advantage of greatly reducing the number of possible subspaces and therefore the required number of signals, as stated in the richness property definition above. Note that

we do not require prior knowledge of the block size  $k_\ell$ , as explained below. We now present our main result.

**Theorem 1 (Uniqueness conditions for blind CS with multiple measurement matrices).** Let  $\mathbf{y}_i \in \mathbb{R}^{m_i}$ , with  $i = 1, \dots, N$ , be a set of observed vectors, obtained through projections of unknown signals  $\mathbf{x}_i \in \mathbb{R}^n$  so that each  $\mathbf{y}_i = \mathbf{A}_i \mathbf{x}_i$  and  $\mathbf{A}_i \in \mathbb{R}^{m_i \times n}$  is a known sensing matrix. Furthermore, let  $\mathbf{x}_i = \mathbf{D} \mathbf{s}_i$  where  $\mathbf{s}_i \in \mathbb{R}^r$  is an unknown vector of coefficients obeying one-block sparsity according to Definition 1 and  $\mathbf{D} \in \mathbb{R}^{n \times r}$  is an unknown dictionary having a block structure such that  $\mathbf{D} = [\mathbf{D}[1] \cdots \mathbf{D}[L]]$ , with  $\mathbf{D}[\ell]^T \mathbf{D}[\ell] = \mathbf{I}$ . Let  $k_\ell$  be the number of columns of block  $\mathbf{D}[\ell]$ . In addition, let  $\omega_\ell$  be the index set of the vectors  $\mathbf{x}_i$  for which the corresponding  $\mathbf{s}_i$  have block  $\mathbf{s}_i[\ell]$  active, i.e., nonzero, and let  $|\omega_\ell|$  denote the number of such vectors. Define  $\tilde{\mathbf{A}} \in \mathbb{R}^{M \times n}$  as the union of the unique rows of all  $\mathbf{A}_i$  for all  $i \in \omega_\ell$ , so that  $\mathbf{Y}_{\omega_\ell} \in \mathbb{R}^{M \times |\omega_\ell|}$  with  $M \geq n$  has columns given by  $\tilde{\mathbf{A}} \mathbf{x}_i$ , for  $i \in \omega_\ell$ , and only a subset of the elements in  $\mathbf{Y}_{\omega_\ell}$  are observed. Define  $M_{1\ell} = \min(M, |\omega_\ell|)$ ,  $M_{2\ell} = \max(M, |\omega_\ell|)$  and let  $\mathbf{Y}_{\omega_\ell} = \mathbf{U} \Sigma \mathbf{V}^T$  be the singular value decomposition of  $\mathbf{Y}_{\omega_\ell}$ . Define  $\mu_\ell = \max(\mu_1^2, \mu_0)$ , where  $\mu_1$  is an upper bound on the absolute value of the entries of  $\mathbf{U} \mathbf{V}^T \sqrt{(M_{1\ell} M_{2\ell})/k_\ell}$  and  $\mu_0$  is an upper bound on the coherence of the row and column spaces of  $\mathbf{Y}_{\omega_\ell}$ .

Then, by solving problem (9) one can exactly recover all the blocks of  $\mathbf{D}$  and the coefficient vectors  $\mathbf{s}_i$  up to the equivalence class presented in Definition 2, with probability at least  $1 - 6 \log(M_2)(M_1 + M_2)^{2-2\beta} - M_2^{2-2\sqrt{\beta}}$  for some  $\beta > 1$ , if the following conditions are met for each  $\ell \in \{1, \dots, L\}$ .

- For all  $i \in \omega_\ell$ ,  $\|\mathbf{s}_i\|_0 = k_\ell < \frac{\sigma(\tilde{\mathbf{A}}\mathbf{D})}{2}$ .
- $|\omega_\ell| > k_\ell$ .
- The vectors  $\mathbf{x}_i, i \in \omega_\ell$ , are non-degenerate, i.e., any

subset of  $k \leq k_\ell$  vectors span a subspace of rank at least  $k$ .

- (iv)  $32\mu_\ell k_\ell(M_{1\ell} + M_{2\ell})\beta \log(2M_{2\ell})$  entries of the matrix  $\mathbf{Y}_{\omega_\ell}$  are observed uniformly at random. The total number of observed entries is, thus,  $\sum_{\ell=1}^L 32\mu_\ell k_\ell(M_{1\ell} + M_{2\ell})\beta \log(2M_{2\ell})$ .

*Proof of Theorem 1:* Condition (i)–(iii) are analogous to the DL uniqueness conditions in Definition 3, which are proven in [1], and are always required even if all  $\mathbf{y}_i$  are fully measured. Condition (i) is adapted to our setting by imposing the spark restriction on  $\tilde{\mathbf{A}}\mathbf{D}$ ; this condition ensures that no two dictionary blocks are linearly dependent, and that the sensing matrices are sufficiently incoherent with respect to the dictionary. This is the case, with high probability, when  $\tilde{\mathbf{A}}$  obeys a random subgaussian or orthobasis construction [6], [9].

Condition (ii) stems from the fact that any  $k_\ell$ -dimensional hyperplane is uniquely determined by  $k_\ell + 1$  vectors, and since a subspace will always contain the origin, only  $k_\ell$  additional vectors are required. We write  $|\omega_\ell| > k_\ell$  instead of  $|\omega_\ell| \geq k_\ell$  because, in the setting of one-block sparsity, there exist  $N = \sum_{\ell=1}^L |\omega_\ell|$  vectors from a union of  $L$  different subspaces, and we need to be able to assign vectors to their respective subspaces.

If condition (iii) holds, which precludes spurious colinearities or coplanarities, then the assignment can be done by the (admittedly impractical) procedure of testing the rank of all possible  $\binom{N}{k_\ell+1}$  matrices constructed by concatenating subsets of  $k_\ell + 1$  column vectors, as assumed in [1]. If the rank of such a matrix is  $k_\ell$ , then all its column vectors belong to the same subspace. Otherwise, the rank will be  $k_\ell + 1$ . It is not necessary to know  $k_\ell$  *a priori*, provided that we test all possible values  $k_\ell = 1, \dots, n - 1$  in the following way. Start by setting  $k_\ell = 1$  and testing all possible subsets containing  $k_\ell + 1 = 2$  signals. If any such subset has rank one, then we have found a (singleton) block. Any other possibility is precluded by the non-degeneracy condition (iii). If multiple subsets have rank one, then test the subspace angles in order to determine how many distinct singleton blocks exist. Record the corresponding blocks and signal assignments and remove the signals involved from further consideration (due to one-block sparsity, each signal belongs to only one block). Iterate this procedure until either: (a) we have exhausted the signals; (b) all  $r$  atoms have been clustered into blocks, or (c) we have reached  $k_\ell = n$ .

We now address Condition (iv). We have  $|\omega_\ell| > k_\ell$  vectors from subspace  $\ell$  and we assume that  $\tilde{\mathbf{A}}$  has rank  $n$  and is, therefore, invertible. As discussed in Section II,  $\mathbf{Y}_{\omega_\ell}$  is an incomplete low-rank matrix. Under conditions studied in [24], the convex program (9) will recover the complete  $\mathbf{Y}_{\omega_\ell}$  with high probability. Therefore, by showing uniqueness of the complete low-rank matrix  $\mathbf{Y}_{\omega_\ell}$ , we can show the uniqueness of the corresponding subspace spanned by  $\mathbf{D}[\ell]$ , which can be found by first solving the optimization (9) and then taking the estimate  $\hat{\mathbf{Y}}_{\omega_\ell}$  and performing the SVD of  $\tilde{\mathbf{A}}^+ \hat{\mathbf{Y}}_{\omega_\ell}$ .

We now invoke Theorem 1.1 in [24] which states that, under the conditions and with the probability specified in our

theorem, the minimizer of problem (9) is unique and equal to the true  $\mathbf{Y}_{\omega_\ell}$ . This concludes the proof. ■

Note that the aforementioned theorem from [24], like preceding results [7], is based on  $\Omega$  having at least one entry for each row and for each column; this is already accounted for in the derivation of the bound on  $|\Omega|$  and associated probability of successful recovery under a uniformly-at-random pattern for the missing data. If there is a fixed  $\mathbf{A}_i = \mathbf{A}$  (with less than  $n$  rows), or if there is any row or column entirely missing from  $\Omega$ , then there is no recovery guarantee. This is the case of standard blind CS, where there is a fixed  $\mathbf{A}$  which is the same for all signals, and thus there are entire rows without observations, as illustrated in Figure 2. Hence, blind CS requires additional constraints on  $\mathbf{D}$ , as explained in [17]. Using multiple  $\mathbf{A}_i$  allows us to avoid those constraints. Also, even when  $M \geq n$ , we are still subsampling because we do not observe all entries of  $\mathbf{Y}_{\omega_\ell}$ .

The limiting case when all blocks are singletons, *i.e.*,  $k_\ell = 1$  for all  $\ell$ , coupled with the one-block sparsity assumption, means that each patch belongs to a straight line in  $\mathbb{R}^n$ , which is a very strong restriction. On the other hand, if all  $k_\ell = 1$  but the one-block sparsity assumption is removed and there are  $K$  active blocks (singleton atoms when  $k_\ell = 1$ ), then we revert to standard sparsity; Theorem 1 still applies, but we would need to be test  $\binom{r}{K}$  possible combinations of atoms, which may be an extremely large number. In fact, the conditions in Theorem 1, with standard sparsity and additionally with all  $\mathbf{A} = \mathbf{I}_{n \times n}$ , reduce to those in DL. One-block sparsity makes Theorem 1 more appealing, since then there are only  $L$  possible combinations instead of  $\binom{r}{K}$ .

Under Theorem 1, recovery is contingent on having the correct clustering given by  $\omega_\ell$ . In the above-presented analysis, it is assumed that we have the computational ability to exhaustively search over all index sets  $\omega_\ell$ . Under the conditions of Theorem 1, this search will achieve the correct assignment of signals to subspaces (clustering), and we need only concern ourselves with vectors that all come from the same subspace. This is done only for the purpose of constructing a proof, following the same strategy as in [1]. Also, like the DL uniqueness result in [1], Theorem 1 is overly pessimistic in the required number of measurements. Specifically, for image inpainting applications, where  $\mathbf{Y}_{\omega_\ell}$  typically has far more columns than rows, the prescribed number of measurements can be excessive (often orders of magnitude larger than the total number of pixels). This reflects a limitation of the current state-of-the-art in matrix completion theory, and we stress that other uniqueness results, such as [1] for the simpler DL setting, suffer from the same problem. Nevertheless, Theorem 1 provides peace of mind by guaranteeing the existence of a unique optimal solution given enough measurements. It remains necessary to address the practical issue of finding a solution with reasonable computational effort, from realistic amounts of unclustered data.

Toward that end, in the next section we propose an algorithm that can estimate the clustering and each of the subspaces without combinatorial search, although it is not guaranteed to reach the globally optimum solution. The algorithm uses alternating least squares, similarly to BSDO, K-SVD and



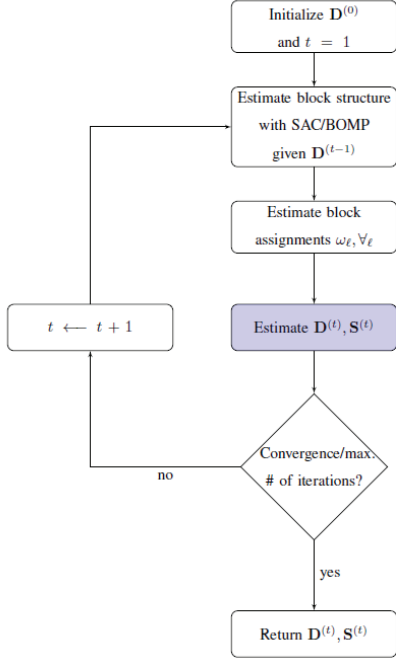


Fig. 3. Block diagram of the overall dictionary learning procedure. The analysis in this paper focuses on the shaded block, where we plug our Algorithm 1 instead of the BK-SVD step in [26].

MOD. It is computationally efficient and enjoys local convergence guarantees given much fewer measurements than those prescribed by Theorem 1. In Section V we show convergence of the algorithm to a local optimum, as long as specified algorithm-specific conditions hold.

#### IV. ALGORITHM

Our algorithm expands on the iterative procedure described in [26]. This procedure alternates between two major steps: (i) inferring the block structure and coefficients using sparse agglomerative clustering (SAC) [26] and block-orthogonal matching pursuit (BOMP) [12], and (ii) updating the dictionary blocks using block-K-SVD (BK-SVD) [26]. The latter step assumes fully measured data. Briefly, SAC progressively merges pairs of blocks  $\ell_1, \ell_2$  for which the index sets  $\omega_{\ell_1}$  and  $\omega_{\ell_2}$  are most similar, up to an upper limit ( $k_{\max}$ ) on the block size. BOMP sequentially selects the blocks that best match the observed signals  $\mathbf{y}_i$  and can be viewed as a generalization of OMP [19], [22] for the block-sparse case. Similarly, BK-SVD is an extension of K-SVD for block-sparsifying dictionaries. A more detailed explanation of these methods may be found in [26] and the references therein.

We follow the procedure in [26], but replace the BK-SVD step by our Algorithm 1 in order to take compressive measurements into account, as the block diagram in Figure 3 illustrates. We employ alternating least-squares minimization to find  $\mathbf{D}$  and  $\mathbf{S}$  in a manner similar to existing methods such as K-SVD or MOD but adapted to the CS setting.

Recall that we are solving the optimization problem (5). Assuming the current estimate of  $\omega_\ell$  and the block structure

are fixed (as estimated in the previous SAC step), then we need to solve, for each  $\ell$ ,

$$\min_{\mathbf{D}[\ell], \mathbf{S}_{\omega_\ell}[\ell]} \sum_{i \in \omega_\ell} \|\mathbf{y}_i - \mathbf{A}_i \mathbf{D}[\ell] \mathbf{s}_i[\ell]\|_2^2 \quad (12)$$

$$\text{s.t.} \quad \mathbf{D}[\ell]^T \mathbf{D}[\ell] = \mathbf{I}.$$

The optimization alternates between  $\mathbf{D}[\ell]$  and  $\mathbf{S}_{\omega_\ell}[\ell]$ . For now we will set the constraint  $\mathbf{D}[\ell]^T \mathbf{D}[\ell] = \mathbf{I}$  aside and focus on the unconstrained problem; we will return to the constraint later. A description of the two main algorithm steps is presented below, followed by the derivation of conditions for convergence.

##### A. Dictionary update

Holding  $\mathbf{s}_i$  fixed, we can analytically find  $\mathbf{D}[\ell]$  using properties of the Kronecker product and the vectorization operator, denoted as  $\text{vec}(\cdot)$ . This is similar to MOD [18] in that the entire dictionary block is updated at once. For a generic matrix  $\mathbf{Z}$ ,  $\text{vec}(\mathbf{Z})$  is the column vector constructed by vertically stacking the columns of  $\mathbf{Z}$ . For simplicity of presentation, and without loss of generality, assume that all  $m_i = m$ . Let  $\tilde{\mathbf{Y}}_{\omega_\ell}$  denote the matrix whose columns are the vectors  $\mathbf{y}_i$ , so that we have  $|\text{vec}(\tilde{\mathbf{Y}}_{\omega_\ell})| = m|\omega_\ell|$ . We rewrite the unconstrained problem over  $\mathbf{D}[\ell]$  as

$$\min_{\mathbf{D}[\ell]} \left\| \text{vec}(\tilde{\mathbf{Y}}_{\omega_\ell}) - \mathbf{B} \text{vec}(\mathbf{D}[\ell]) \right\|_2^2, \quad (13)$$

where  $\mathbf{B}$  is defined as

$$\mathbf{B} = \begin{bmatrix} \mathbf{s}_{i_1}[\ell]^T \otimes \mathbf{A}_{i_1} \\ \vdots \\ \mathbf{s}_{i_{|\omega_\ell|}}[\ell]^T \otimes \mathbf{A}_{i_{|\omega_\ell|}} \end{bmatrix}, \quad (14)$$

with  $\otimes$  denoting the Kronecker product. This is a least-squares optimization problem for a standard system of linear equations. The solution

$$\text{vec}(\mathbf{D}[\ell]) = \mathbf{B}^+ \text{vec}(\tilde{\mathbf{Y}}_{\omega_\ell}) \quad (15)$$

is well-defined and unique (recall that we are holding the  $\mathbf{s}_i$ , and therefore  $\mathbf{B}$ , fixed) as long as  $\mathbf{B}$  has rank equal to  $k_\ell n$ , the number of unknowns in  $\text{vec}(\mathbf{D}[\ell])$ .

##### B. Coefficient update

After obtaining  $\mathbf{D}[\ell]$  for the current iteration, the coefficient vectors  $\mathbf{s}_i[\ell]$  are found analytically by least squares as

$$\mathbf{s}_i[\ell] = (\mathbf{D}^T[\ell] \mathbf{A}_i^T \mathbf{A}_i \mathbf{D}[\ell])^{-1} (\mathbf{D}^T[\ell] \mathbf{A}_i^T \mathbf{x}_i). \quad (16)$$

The factor  $(\mathbf{D}^T[\ell] \mathbf{A}_i^T \mathbf{A}_i \mathbf{D}[\ell])^{-1}$  is invertible under conditions specified in our convergence result below.

##### C. Orthogonality constraint

The orthogonality condition  $\mathbf{D}[\ell]^T \mathbf{D}[\ell] = \mathbf{I}$  is currently imposed *a posteriori* by performing a SVD decomposition of  $\mathbf{D}[\ell]$  and applying the resulting unitary rotation matrix to the  $\mathbf{s}_i[\ell]$ . This is equivalent to Gram-Schmidt orthogonalization, and will not change the locations of the nonzero elements of the  $\mathbf{s}_i$ , although the blocks  $\mathbf{D}[\ell]$  will not have orthogonal

columns during the iterative procedure. This procedure has worked well in practice, and we therefore do not pursue a more direct constrained minimization.

In addition to the proposed Algorithm 1, we have also implemented and tested an alternative method where, instead of alternating least squares, we perform convex optimization of  $\mathbf{D}$  and  $\mathbf{S}$  jointly (given the estimate of  $\omega_\ell$  provided by the current SAC step) by completing matrix  $\mathbf{Y}_{\omega_\ell}$ , which is closer in spirit to the ideas underlying the uniqueness conditions in Theorem 1. However, existing general-purpose convex optimization packages do not readily scale up to matrices containing millions of entries as in our inpainting experiments, as noted in [3]. Using mixed nuclear/Frobenius norm minimization via the singular value thresholding (SVT) approach proposed in [3], we have attained reconstruction performance comparable to our Algorithm 1, but at much higher computational cost. We have also found that successful convergence with this alternative SVT-based approach depends on careful tuning of step-size and regularization parameters. In contrast, our alternating least squares method does not have any tuning parameters (other than the maximum block size  $k_{\max}$ ) and is computationally far less demanding.

---

**Algorithm 1** – Joint estimation of  $\mathbf{D}[\ell]$  and  $\mathbf{S}_{\omega_\ell}[\ell]$  via alternating minimization

---

**Initialization:** Use the estimates of  $\omega_\ell$ ,  $\mathbf{S}_{\omega_\ell}$  and the block structure of  $\mathbf{D}$  from the preceding SAC/BOMP step

**for all**  $\ell$  **do**

- Form the matrix  $\mathbf{B} = \begin{bmatrix} \mathbf{s}_{i_1}[\ell]^T \otimes \mathbf{A}_{i_1} \\ \vdots \\ \mathbf{s}_{i_{|\omega_\ell|}}[\ell]^T \otimes \mathbf{A}_{i_{|\omega_\ell|}} \end{bmatrix}$
- Update block  $\ell$  of the dictionary by computing  $\text{vec}(\mathbf{D}[\ell]) = \mathbf{B}^+ \text{vec}(\tilde{\mathbf{Y}}_{\omega_\ell})$
- Orthogonalize  $\mathbf{D}[\ell]$
- Update the coefficients:  $\mathbf{s}_i[\ell] = (\mathbf{D}^T[\ell] \mathbf{A}_i^T \mathbf{A}_i \mathbf{D}[\ell])^{-1} (\mathbf{D}^T[\ell] \mathbf{A}_i^T \mathbf{x}_i), \forall i \in \omega_\ell$

**end for**

---

## V. CONVERGENCE RESULT

The following result establishes convergence conditions for our proposed alternating least-squares algorithm.

**Proposition 1 (Convergence of Algorithm 1).** *Let  $\mathbf{y}_i \in \mathbb{R}^{m_i}$ , with  $i = 1, \dots, N$  be a set of vectors that satisfy the conditions stated in Theorem 1. Let  $\mathbf{A}_i \in \mathbb{R}^{m_i \times N}$  be sensing matrices and  $\mathbf{x}_i \in \mathbb{R}^n$  signals such that  $\mathbf{y}_i = \mathbf{A}_i \mathbf{x}_i$ . Assume each signal is of the form  $\mathbf{x}_i = \mathbf{D} \mathbf{s}_i$ , with  $\mathbf{D}$  a dictionary comprised of blocks  $\mathbf{D}[1], \dots, \mathbf{D}[L]$ , where each block has  $k_\ell$  atoms, and  $\mathbf{s}_i \in \mathbb{R}^r$  a one-block-sparse vector. Let  $\omega_\ell$  be the index set of vectors  $\mathbf{x}_i$  from dictionary block  $\ell$ , and let  $\mathbf{Y}_{\omega_\ell} \in \mathbb{R}^{M \times |\omega_\ell|}$  be an incomplete data matrix such that  $P_\Omega(\mathbf{Y}_{\omega_\ell}) = P_\Omega(\tilde{\mathbf{A}} \mathbf{D}[\ell] \mathbf{S}_{\omega_\ell}[\ell])$ , with  $\Omega$  the locations of observed entries, and  $\tilde{\mathbf{A}} \in \mathbb{R}^{M \times n}$  the union of the rows of the  $\mathbf{A}_i$ . Then, the alternating minimization procedure described*

*in Section IV will converge to a local minimum of (4), with probability specified below, if the following conditions hold:*

- (i) *For all  $\ell$  and all  $i \in \omega_\ell$ ,  $\|\mathbf{s}_i\|_0 = k_\ell < \frac{\sigma(\tilde{\mathbf{A}} \mathbf{D})}{2}$ .*
- (ii) *For the subset of vectors associated with each block  $\ell$ , we have  $|\omega_\ell| \geq n$ .*
- (iii) *The total number of observed values,  $|\Omega|$ , is  $O(k_\ell n)$  if the elements of all  $\mathbf{A}_i$  are i.i.d Gaussian, and  $O(k_\ell n \log n)$  if the rows of  $\mathbf{A}_i$  are selected uniformly at random from  $\tilde{\mathbf{A}}$  and  $\text{rank}(\tilde{\mathbf{A}}) = n$ .*
- (iv) *Each vector  $\mathbf{y}_i$  has  $m_i \geq k_\ell$  measured values.*

*The probability of convergence, for each block  $\ell$ , is equal to one with i.i.d. Gaussian  $\mathbf{A}_i$ , and equal to  $1 - n^{-\beta+1}$  for random selection of rows from  $\tilde{\mathbf{A}}$ , where  $\beta$  is the constant such that  $|\Omega| = \beta n \log n$ .*

*Proof of Proposition 1:* Since condition (i) is the same as in Theorem 1, we focus on conditions (ii)–(iv). As mentioned above, the analytic solution (15) is well-defined and unique as long as  $\mathbf{B}$  has rank equal to  $k_\ell n$ . Examining (13) and (14) we can see that, if the  $\mathbf{s}_i[\ell]$  are non-degenerate (i.e.,  $\text{rank}(\mathbf{S}_{\omega_\ell}[\ell]) = k_\ell$ , with  $\mathbf{S}_{\omega_\ell}[\ell]$  the subset of nonzero rows of  $\mathbf{S}_{\omega_\ell}$ ), then the rank condition is equivalent to having the number of observed elements  $|\text{vec}(\tilde{\mathbf{Y}}_{\omega_\ell})| = |\Omega| \geq k_\ell n$  and also

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{A}_{i_1} \\ \vdots \\ \mathbf{A}_{i_{|\omega_\ell|}} \end{bmatrix}, \quad (17)$$

that is, the vertical concatenation of all  $\mathbf{A}_i$ , having rank  $n$ . This is due to the fact that  $\text{rank}(\mathbf{s}_i \otimes \mathbf{A}_i) = \text{rank}(\mathbf{s}_i) \times \text{rank}(\mathbf{A}_i), \forall i$ . The number of observed values needed to ensure this rank condition depends on the probabilistic mechanism that generates the  $\mathbf{A}_i$ . We will analyze two such mechanisms: random Gaussian and random subset of projections. Without loss of generality, assume all  $m_i = m$  (we can always treat  $m$  as a lower limit on the  $m_i$ ).

**Random Gaussian:** We assume that all  $\mathbf{A}_i$  have elements independently drawn from a Gaussian distribution. Then, with probability one, any subset of size  $n$  of the  $m|\omega_\ell|$  rows of  $\mathbf{\Gamma}$  are linearly independent. Therefore, the condition  $m|\omega_\ell| \geq n$  ensures  $\text{rank}(\mathbf{\Gamma}) = n$ ; since we need  $m|\omega_\ell| \geq k_\ell n$  (and therefore  $|\omega_\ell| > n$ ) in order to have enough observed entries anyway, the latter condition ensures that the rank of  $\mathbf{\Gamma}$  is adequate. Note that this reasoning assumes that the  $\mathbf{A}_i$  do not share rows; if they do share rows, e.g., in the case when there is a finite pool of random-Gaussian projections and the rows of  $\mathbf{A}_i$  are subsets from the pool, then we must use the following result instead.

**Random subset of projections:** We analyze the case for which the rows of each sensing matrix  $\mathbf{A}_i$  are randomly drawn from a pool of linearly independent rows, the union of which constitutes  $\tilde{\mathbf{A}}$ . The typical example of this situation is when the rows of  $\tilde{\mathbf{A}}$  form a random orthonormal basis, although here orthonormality is not required (linear independence suffices). This includes the case when  $\tilde{\mathbf{A}}$  is the identity matrix, which is most relevant to the inpainting/interpolation problem. Note that the problem can be treated as an instance of the classic Coupon Collector problem (see, e.g., [15]). There are  $m|\omega_\ell|$  rows (the



number of rounds our “collector” draws a randomly chosen row, or “coupon”) that have to span a space of  $n$  dimensions (the different types of “coupons” we need to collect). Denote as  $Z_i^{m|\omega_\ell|}$  the event that no coupons of type  $i$  have been drawn after  $m|\omega_\ell|$  rounds. It is easy to see that

$$P[Z_i^{m|\omega_\ell|}] \leq \left(1 - \frac{1}{n}\right)^{m|\omega_\ell|}, \quad (18)$$

where the inequality comes from our ignoring, for simplicity, the fact that the rows are drawn without replacement within each sensing matrix (stated otherwise, there are no repeated rows within each  $\mathbf{A}_i$ ). Then, by the union bound,

$$P\left[\bigcup_{i=1,\dots,n} Z_i^{m|\omega_\ell|}\right] \leq n \left(1 - \frac{1}{n}\right)^{m|\omega_\ell|}. \quad (19)$$

The left-hand side is precisely the probability that  $\mathbf{\Gamma}$  is not rank  $n$ . To look at how this bound scales relative to  $n$ , we can apply the inequality  $1 - x \leq e^{-x}$  to obtain  $\left(1 - \frac{1}{n}\right)^{m|\omega_\ell|} \leq e^{-\frac{m|\omega_\ell|}{n}}$ . Then, consider  $m|\omega_\ell| = \beta n \log n$ , with  $\beta$  constant, and define  $T$  to be the minimum value of  $m|\omega_\ell|$  before we achieve full rank, so that  $P[T > m|\omega_\ell|] = P[T > \beta n \log n] = P\left[\bigcup_{i=1,\dots,n} Z_i^{m|\omega_\ell|}\right]$ . We have

$$P[T > \beta n \log n] \leq n \left(1 - \frac{1}{n}\right)^{m|\omega_\ell|} \leq n e^{-\frac{m|\omega_\ell|}{n}} = n e^{-\beta n \log n} \quad (20)$$

Therefore, we need  $O(n \log n)$  observed values in order to achieve rank  $n$  with arbitrarily high probability. Hence, we need  $O(n k_\ell \log n)$  for solving (13) for each block. Although we do not impose the orthogonality constraint directly in the minimization, any solution of (13) can be orthonormalized and thus become a solution of (12) *without changing the subspace* defined by  $\mathbf{D}[\ell]$ .

Concerning the coefficient update expression (16), the issue of invertibility comes up due to the term  $(\mathbf{D}^T[\ell] \mathbf{A}_i^T \mathbf{A}_i \mathbf{D}[\ell])^{-1}$ . Note that the matrix being inverted has size  $k_\ell \times k_\ell$ . The rank of  $\mathbf{A}_i^T \mathbf{A}_i$  is equal to  $m_i$ , the number of rows of  $\mathbf{A}_i$ . For invertibility of the above expression, we need to have  $k_\ell \leq m_i$ , so that there is no loss of rank. This means that our method requires a lower bound on  $m_i$ .

We conclude the proof by invoking, as in [1], the fact that all of the steps of this alternating minimization method are optimal under the above assumptions, and hence cannot increase the value of the objective function. Therefore, since the objective function is bounded from below by zero, the algorithm will converge. ■

We have proven convergence in the objective function, which is a weak form of convergence in the sense that the objective function will attain a minimum but the estimates might not reach a stopping point. However, convergence in the objective function is still a useful result, and it is the same guarantee as K-SVD and MOD. In practice, the above result ensures that the required number of signals scales linearly with  $k_\ell$  and also linearly (up to a log factor in the case of measurements using an orthobasis ensemble) with signal dimensionality  $n$ . These are more favorable bounds than those

in Theorem 1 for uniqueness. However, this is not surprising since Theorem 1 is based on matrix completion theory, which establishes slightly pessimistic sufficient conditions for strong unique global recovery guarantees, while Proposition 1 only establishes convergence of our algorithm to a local minimum. It is also interesting that, in the case of random subsets of projections, the  $O(k_\ell n \log n)$  requirement matches the information-theoretic limit established in [7] for matrix completion. This similarity is due to the fact that, as in [7] and other work, we make use of the Coupon Collector model. Regarding the lower bound on  $m_i$ , since we typically expect the block size to be small, the condition  $m_i \geq k_\ell$  is relatively mild and should only be of concern for extremely small measurement fractions.

## VI. EXPERIMENTAL RESULTS

The algorithm proposed in Section IV is validated by inpainting the well-known “Barbara” ( $512 \times 512$  pixels) and “house” ( $256 \times 256$  pixels) images, for varying percentages of observed pixels (these two examples are representative of many others we have considered). The images are processed in  $8 \times 8$  overlapping patches, treated as vectors of dimension  $n = 64$ . In Figures 4 and 5 we present the original images and the test versions, with 25%, 50% and 75% of the pixel values observed (selected uniformly at random). In all experiments, the total number of dictionary elements is set to  $r = 256$ . Figures 6 and 7 show the inpainting results achieved by our algorithm from 50% observed pixels, using maximum block sizes  $k_{\max} = 4$  and  $k_{\max} = 8$ . The peak signal-to-noise ratios (PSNR) for this case are shown in Table II.

TABLE II  
PEAK SIGNAL-TO-NOISE RATIOS (PSNR) IN INPAINTING TASKS WITH  
50% OBSERVED PIXELS

	$k_{\max} = 4$	$k_{\max} = 8$
“Barbara”	27.68 dB	27.93 dB
“House”	31.80 dB	32.03 dB

The PSNR values for varying percentages of observed pixels are plotted in Figure 8, for  $k_{\max}$  equal to four and eight. Results are averaged over ten runs with different random locations for the missing pixels, with error bars showing the one standard deviation interval. These experiments are intended to validate our method, rather than claiming outperformance of the state-of-the-art in image inpainting and interpolation. Our performance is comparable to that of the algorithm described in [31], which we have used on the same images. However, our model and algorithm are significantly simpler and, unlike [31], our algorithm has convergence guarantees. Also, while the PSNR achieved in our experiments is lower than in the state-of-the-art approach in [30], the results are not directly comparable due to the fact that additional structure is assumed in [30], including fast decay of the singular values within each block (which is analogous to internal sparsity).

Like other authors (e.g., [16]), we have observed a phase transition phenomenon where the reconstruction performance falls sharply for measurement percentages under a rank-dependent threshold (near 40% observed pixels in our case).

This is shown in Figure 9, which was obtained using as input subsets of pixels from a union-of-subspaces approximation of the “Barbara” image for block sizes four and eight, thus matching our model exactly. We plot the empirical frequency of successful reconstructions over ten realizations of the missing pixel locations, for varying measurement percentages. The reconstruction is deemed successful when the PSNR exceeds 40 dB.

In addition, the learned dictionaries for  $k_{\max} = 4$  and  $k_{\max} = 8$  are depicted in Figures 10 and 11. It is possible to observe that the atoms within each dictionary block are more similar to each other than those in different blocks. This is expected and desired clustering behavior, as one would like the blocks to reflect different image properties. This behavior is further illustrated in Figure 12, where the index of the block used by each patch is represented by a corresponding color (to avoid clutter, this is shown for the most frequently used dictionary blocks). It is apparent that patches with similar texture tend to share blocks.

In all cases tested, convergence occurs after a maximum of ten iterations. A non-optimized MATLAB implementation of Algorithm 1, including the SAC/BOMP steps, requires approximately 12 hours for inpainting the full  $512 \times 512$  “Barbara” image (worst case) on a computer with CPU clock frequency of 2.52 GHz. This computation time is similar to the BSDO algorithm described in [26]. We employ overlapping patches in order to avoid block artifacts, leading to a total of  $N = 255,025$  vectors being processed for the “Barbara” image, with  $N = 62,001$  for the “house” image. Note that the overlapping procedure follows the same “non-overlapping sensing with overlapping reconstruction” strategy as [29], [31]; this procedure employs a distinct sensing matrix per *non-overlapping* patch, with the image reconstruction performed by averaging the *overlapping* estimated patches.

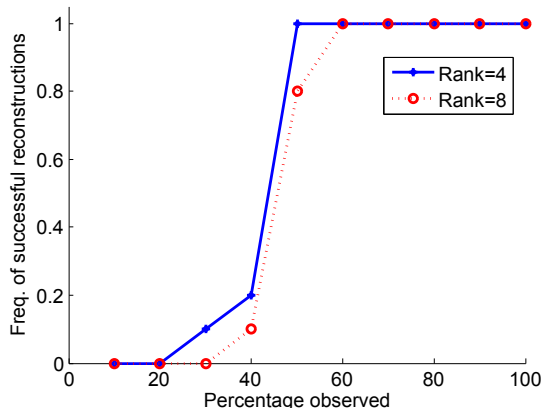


Fig. 9. Illustration of the phase transition phenomenon, using as input subsets of pixels from an approximation of the “Barbara” image that exactly obeys the union-of-subspaces model, with block size (rank) four and eight. The plot shows the normalized frequency of successful reconstructions over ten realizations of missing pixel locations, for varying percentages of observed pixels. We declare a successful reconstruction when the PSNR of the estimate is greater than 40 dB.

## VII. CONCLUSION

We have proposed a framework for simultaneous estimation of one-block-sparse signals and the corresponding dictionary, based on compressive measurements. Multiple sensing matrices are employed, which allows use of low-rank matrix completion results to guarantee unique recovery (up to a specified equivalence class) for a sufficiently large number of signals and measurements per signal; bounds are derived for the number of measurements. The assumption of one-block sparsity is related to the union-of-subspaces signal model and to the well-known mixture of factor analyzers (MFA) statistical framework. Existing results for the related problems of dictionary learning (DL) and block-sparse dictionary optimization (BSDO) are extended, due to the analysis of compressive measurements, and blind compressed sensing (blind CS) is also extended by considering a broader class of dictionaries.

Additionally, a practical algorithm based on alternating least-squares minimization has been proposed. The algorithm is related to BSDO, and has been shown to converge to a local optimum under mild conditions. We observe encouraging performance on image inpainting tasks without the need for parameter tuning or careful initialization. An avenue for further research is the treatment of observation noise, which constitutes a natural extension of our work, as well as extension of the theory beyond one-block sparsity.

## REFERENCES

- [1] M. Aharon, M. Elad, and A.M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and Its Applications*, 416(1):48–67, 2006.
- [2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.
- [3] J.F. Cai, E.J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *preprint*, 2008.
- [4] E.J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- [5] E.J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- [6] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- [7] E.J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- [8] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin. Compressive Sensing on Manifolds Using a Nonparametric Mixture of Factor Analyzers: Algorithm and Performance Bounds. 2009.
- [9] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- [10] J.M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *Image Processing, IEEE Transactions on*, 18(7):1395–1408, 2009.
- [11] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- [12] Y.C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *Signal Processing, IEEE Transactions on*, 58(6):3042–3054, 2010.
- [13] Y.C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, 2009.
- [14] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797, 2009.



Fig. 4. Top: original  $512 \times 512$  “Barbara” image. Bottom, left to right: test versions with 25%, 50% and 75% observed pixel values (the remainder are removed).

- [15] W. Feller. *An introduction to probability theory and its applications*, Vol. 1. 1968.
- [16] Quan Geng, Huan Wang, and John Wright. On the local correctness of  $L^1$  minimization for dictionary learning. *CoRR*, abs/1101.5672, 2011.
- [17] S. Gleichman and Y. C. Eldar. Blind Compressed Sensing. *submitted to IEEE Transactions on Information Theory; CCIT Report 759 Feb. 2010, EE Pub No. 1716, EE Dept., Technion - Israel Institute of Technology; [Online] arXiv 1002.2586*, 2010.
- [18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.
- [19] S.G. Mallat and Z. Zhang. Matching pursuits with time–frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.
- [20] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.
- [21] A. Mooijjaart and J.J.F. Commandeur. A general solution of the weighted orthonormal Procrustes problem. *Psychometrika*, 55(4):657–663, 1990.
- [22] Y.C. Pati, R. Rezaifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993.
- [23] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3508, 2010.
- [24] B. Recht. A simpler approach to matrix completion. *CoRR*, abs/0910.0651, 2009.
- [25] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *Submitted to SIAM Review*, 2007.
- [26] K. Rosenblum, L. Zelnik-Manor, and Y.C. Eldar. Dictionary Optimization for Block-Sparse Representations. *Arxiv preprint arXiv:1005.0202*, 2010.
- [27] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.
- [28] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.
- [29] G. Yu and G. Sapiro. Statistical Compressive Sensing of Gaussian Mixture Models. *Arxiv preprint arXiv:1010.4314*, 2010.
- [30] G. Yu, G. Sapiro, and S. Mallat. Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity. *Arxiv preprint arXiv:1006.3056*, 2010.
- [31] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric bayesian dictionary learning for sparse image representations. In *Neural Information Processing Systems (NIPS)*, 2009.

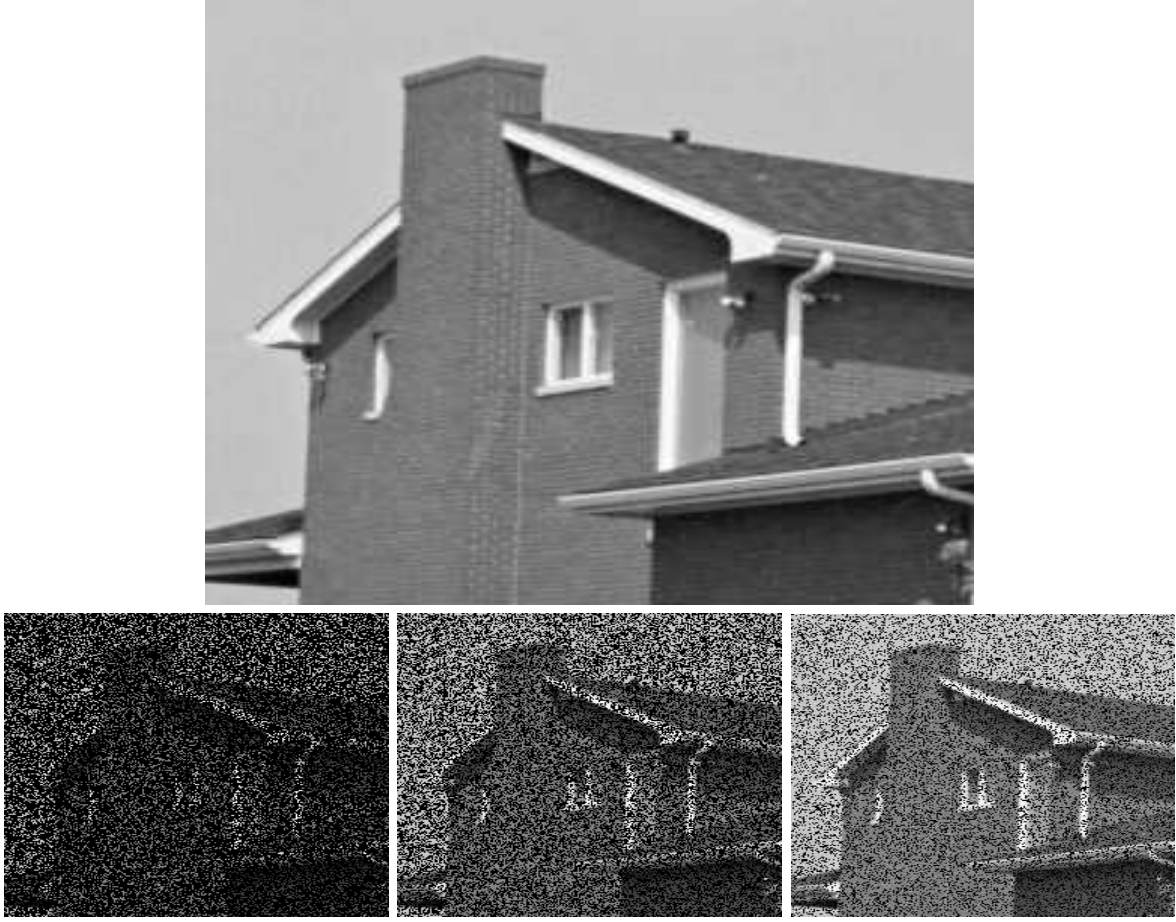


Fig. 5. Top: original  $256 \times 256$  “house” image. Bottom, left to right: test versions with 25%, 50% and 75% observed pixel values (the remainder are removed).



Fig. 6. Inpainted  $512 \times 512$  “Barbara” image with 50% observed pixels. The peak signal-to-noise ratio (PSNR) of this estimate is 27.93 dB. Maximum block size ( $k_{\max}$ ) is eight and the total of dictionary elements is set to  $r = 256$ , divided in  $L = 32$  blocks. Image patches have size  $8 \times 8$  and are treated as vectors of dimension  $n = 64$ . As we employ overlapping patches, the total number of vectors is  $N = 255,025$ .



Fig. 7. Inpainted  $256 \times 256$  “House” image with 50% observed pixels. The peak signal-to-noise ratio (PSNR) of this estimate is 31.80 dB. Maximum block size ( $k_{\max}$ ) is four and the total of dictionary elements is set to  $r = 256$ , divided in  $L = 64$  blocks. Image patches have size  $8 \times 8$  and are treated as vectors of dimension  $n = 64$ . As we employ overlapping patches, the total number of vectors is  $N = 62,001$ .

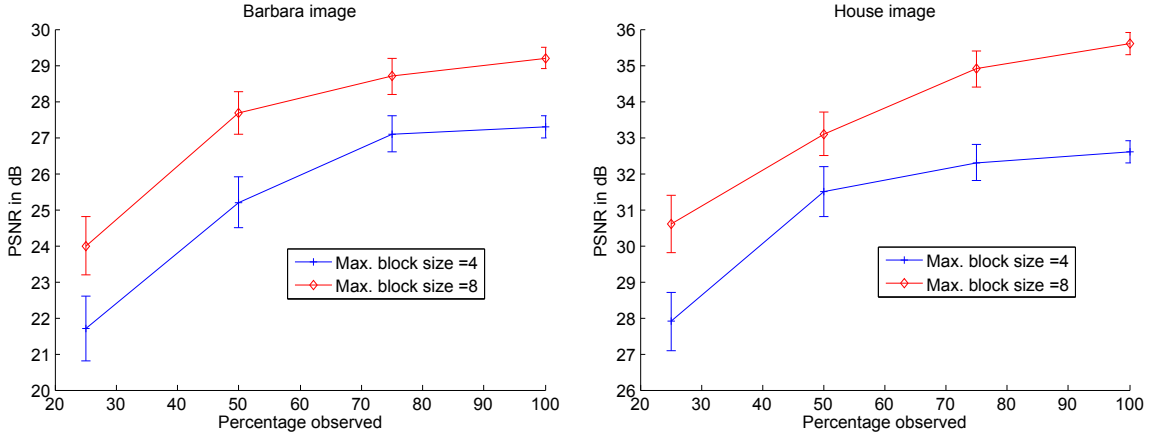


Fig. 8. Peak signal-to-noise ratio (PSNR) for inpainting the “Barbara” (left) and “house” (right) images with 25%, 50%, 75% and 100% observed pixel values, using maximum block size ( $k_{\max}$ ) of four and eight. Results are averaged over ten realizations of the randomly missing pixel locations. The error bars depict one standard deviation. Using higher  $k_{\max}$  yields better PSNR (but also a less parsimonious model).

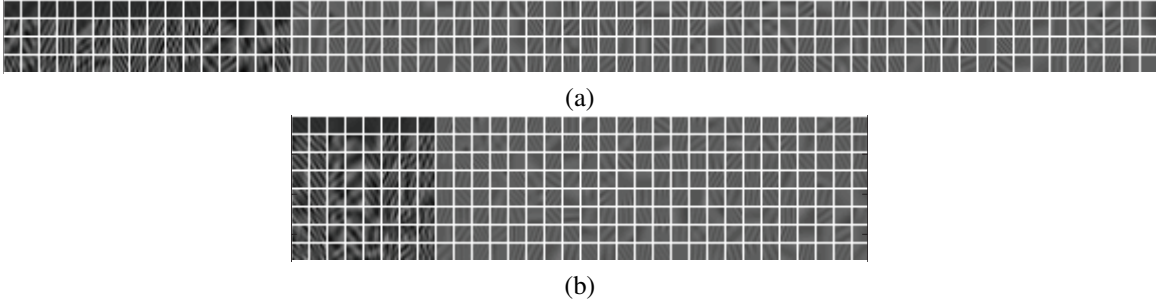


Fig. 10. Learned dictionaries with Algorithm 1 for the “Barbara” image using  $k_{\max} = 4$  (a) and  $k_{\max} = 8$  (b). Each atom is shown as a  $8 \times 8$  image. The atoms are arranged in a  $k_{\max} \times L$  matrix, with all atoms in each column belonging to the same dictionary block.

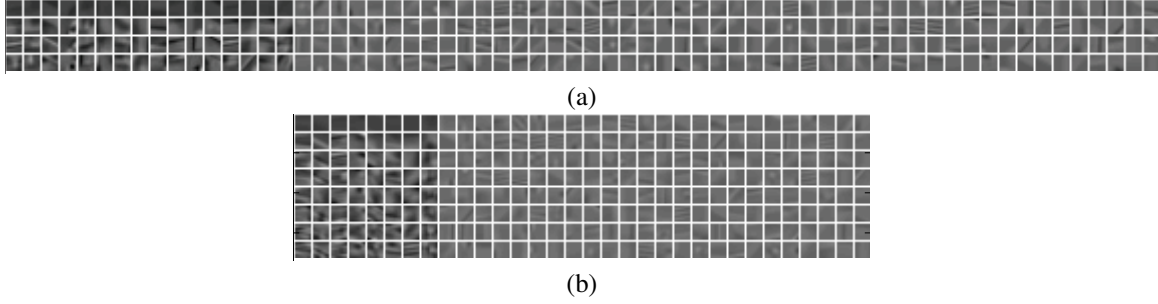


Fig. 11. Learned dictionaries for the “house” image using  $k_{\max} = 4$  (a) and  $k_{\max} = 8$  (b). Each atom is shown as a  $8 \times 8$  image. The atoms are arranged in a  $k_{\max} \times L$  matrix, with all atoms in each column belonging to the same dictionary block.

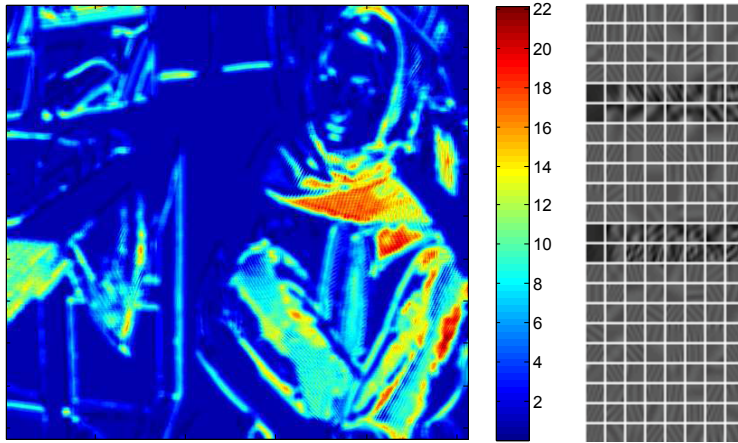


Fig. 12. Dictionary block usage for the “Barbara” image with  $k_{\max} = 8$ . The colors represent the index of the block used by each patch, for a subset containing the 22 most frequently used blocks, out of a total of 32. The subset is shown to the right, aligned with the indices of the colorbar, with each block corresponding of a row of subimages (atoms). It is apparent that patches with similar textures (*e.g.*, the pants, portions of the scarf and table cloth) tend to share blocks (best viewed in color).